Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

(2)

REPORT DOCUMENTATION PAGE

AD-A210 541

JLE

1b RESTRICTIVE MARKINGS: None

3 DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release and sale. Distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S): ONR Technical Report No. 13

5. MONITORING ORGANIZATION REPORT NUMBER(S):

6a. NAME OF PERFORMING ORGANIZATION: University of Utah

6b. OFFICE SYMBOL (*If applicable*):

7a NAME OF MONITORING ORGANIZATION:

6c. ADDRESS (*City, State, and ZIP Code*): Department of Chemistry
Henry Eyring Building
Salt Lake City, UT 84112

7b ADDRESS (*City, State, and ZIP Code*):

8a. NAME OF FUNDING / SPONSORING ORGANIZATION: Office of Naval Research

8b OFFICE SYMBOL (*If applicable*):

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-89-J-1412

8c. ADDRESS (*City, State, and ZIP Code*): Chemistry Program, Code 1113
800 N. Quincy Street
Arlington, VA 22217

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE (*Include Security Classification*): **Advances in Regression: Use of Models in Spectroscopic Data Analysis**

12 PERSONAL AUTHOR(S): J. M. Harris, S. D. Frans, P. E. Poston, and A. L. Wong

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Technical	FROM 7/88 TO 7/89	July 1, 1989	39

16. SUPPLEMENTARY NOTATION:

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*): Method of maximum likelihood, multidimensional data analysis.

19 ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Attached.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION: Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Robert J. Nowak

22b TELEPHONE (*Include Area Code*): (202) 696-4410

22c. OFFICE SYMBOL:

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

OFFICE OF NAVAL RESEARCH

Grant No: N00014-89-J-1412

R&T Code 413a005---03

Technical Report No. 13

Advances in Regression: Use of Models in Spectroscopic Data Analysis

Prepared for publication in Computer-Enhanced Analytical Spectroscopy

by

J. M. Harris, S. D. Frans, P. E. Poston, and A. L. Wong

Department of Chemistry
University of Utah
Salt Lake City, UT 84112

July 1, 1989

ADVANCES IN REGRESSION: USE OF MODELS IN SPECTROSCOPIC DATA ANALYSIS

A Chapter Prepared for "Computer-Enhanced Analytical Spectroscopy" by

J. M. Harris, S. D. Frans, P. E. Poston, and A. L. Wong

Department of Chemistry
University of Utah
Salt Lake City, UT 84112

Advances in Regression: Use of Models in Spectroscopic Data Analysis

CONTENTS

1. INTRODUCTION

Redundant or correlated behavior of spectroscopic data can be costly to the number of degrees of freedom of a measurement and thereby limit the information content of a spectroscopic method of analysis.[1,2] Correlations can take on many forms, including the band shapes of spectroscopic lines, reproducible patterns of spectral features and time-dependent signals, or the predictable variation of component concentrations in hyphenated spectroscopic methods.[3] While such correlated behavior limits the ultimate informing capabilities of a spectroscopic technique, our knowledge of this behavior represents valuable prior information which may be brought to bear on the analysis of the data. Expectations of correlations, in the form of a model, can be used to selectively filter out random fluctuations in data and extract meaningful information in the presence of noise.

Regression analysis[4] or the method of least squares is one of the oldest methods of statistical data analysis, having been first developed in the early 1800's by Gauss and independently by Legendre. The method provides a general approach to extracting underlying relationships from data, including the parameters which describe the relationship between points and the uncertainties in those parameters. Regression methods have their roots in the method of maximum likelihood[5] which assures that the parameter estimates are unbiased and efficient. In this chapter, regression analysis of spectroscopic data will be presented, with emphasis on using models to describe the correlations which are expected in the data, on proper weighting of observations, and on determining uncertainties in estimated parameters. While this approach is particularly powerful for multidimensional, hyphenated spectroscopic methods (time-resolved fluorescence, GC-MS, LC-UV, etc.), the theory of regression methods will first

be developed with examples from measurements of lower dimensionality. The basis of regression methods for multidimensional data in the simple statistics of estimating a mean and standard deviation provides an intuitive basis for understanding more powerful analysis procedures, while extending our background simple statistics into methods for manipulating spectroscopic data.

2. ANALYSIS OF ZERO-DIMENSIONAL SPECTROSCOPIC DATA (NUMBERS)

2.1. Method of Maximum Likelihood

The simplest of spectroscopic measurements provide an outcome which is only a number, the variation of which with an independent variable such as wavelength is not considered. An example of such a measurement is a "colorimetric" analysis, where a sample is reacted with a chromogenic reagent and the absorbance of the product is determined at a single wavelength. Let us assume that we have made a series of N such measurements, x_i, drawn from population of described by a normal distribution having a mean, μ, and a standard deviation, σ_i, which can vary with measurement. Given these results, we wish to determine the "maximum likelihood" estimate of the mean, $\hat{m}$, that is an estimate of the mean of the underlying distribution which would maximize the probability that we observed these results.[5] The probability of observing a series of events is the product of probabilities for observing the individual events; thus, the probability of having observed the N measurements, x_i, is:

$$P_N = \prod_{i=1}^{N} P_i(x_i, m, \sigma_i) \qquad (1)$$

where P_i is normally distributed:

$$P_i = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma_i}\right)^2\right] \quad (2)$$

Taking the product of Gaussian probabilities as a summation within the exponential gives the following expression for the probability of having observed the N results:

$$P_N = \left[\prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - m}{\sigma_i}\right)^2\right] \quad (3)$$

In order to maximize P_N, we minimize the argument of the exponent with respect to $\hat{m}$:

$$\frac{\partial}{\partial m}\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - m}{\sigma_i}\right)^2\right] = 0 = \sum_{i=1}^{N}\left(\frac{x_i - m}{\sigma_i^2}\right) \quad (4)$$

Note that Equation 4 is a "least squares" expression which minimizes the squared deviations between the mean estimate, $\hat{m}$, and the observed data, x_i, weighted by the inverse of the variance of the observations, σ_i^2. Solving Equation 4 for $\hat{m}$ gives the estimate of the mean of the underlying distribution which maximizes the probability that we observed the particular series of N measurements:

$$m \equiv \hat{x} = \left[\sum_{i=1}^{N}\left(x_i/\sigma_i^2\right)\right] \Big/ \sum_{i=1}^{N}\left(1/\sigma_i^2\right) \quad (5)$$

If the uncertainty of each of the measurements is constant, $\sigma_i = \sigma$, then $1/\sigma^2$ can be factored out of the summation, and the maximum likelihood estimate of

the mean given by Equation 5 is simply the average of the N measurements.

The uncertainty of the mean estimate determined by Equation 5 can be found from a propagation of errors[4,5] applied to this expression:

$$\sigma_m^2 = \sum_{i=1}^{N} \left(\frac{\partial m}{\partial x_i}\right)^2 \sigma_i^2 = 1 \Big/ \sum_{i=1}^{N} \left(1/\sigma_i\right)^2 \qquad (6)$$

Again if the uncertainty of each measurement is constant, $\sigma_i = \sigma$, then Equation 6 predicts that the variance of the average of N measurements is 1/N times smaller than the variance of the individual measurements.

2.2. Maximum Likelihood Quantitative Estimates for Peaks

A common goal in analytical spectroscopy is to estimate the concentration of a sample which is responsible for an observed peak, which rises from the baseline as a function of wavelength, frequency, or time. For such data, a number of strategies may be implemented to estimate the sample concentration including measurements of peak height or peak area. The maximum likelihood method, developed above, provides an optimum method of data analysis for such cases.[6] To apply this method, consider N measurements of a spectroscopic signal across a peak, $z_i = cg_i + e_i$, where c is the true sample concentration, g_i is a model peak shape function, and e_i is the error in the measured signal. Under these conditions, each data point provides a measure of sample concentration,

$$c_i = z_i/g_i \qquad (7)$$

the uncertainty of which depends on nature of the errors, e_i, in the measured signal.

If the noise or error in the signal is constant independent of signal

amplitude, $\sigma_{z_i} = \sigma_z$, then the standard deviation of the concentration estimate varies inversely with the peak shape function, $\sigma_{c_i} = \sigma_z/g_i$. Substituting this uncertainty relation and Equation 7 into Equation 5 provides a maximum likelihood estimate of the sample concentration, $\hat{c}$,

$$\hat{c} = \left[\sum_i z_i g_i\right] / \sum_i g_i^2 \qquad (8)$$

This result corresponds to calculating the zero-displacement value of the cross-correlation between the signal and the shape function and is identical to a "matched filter" estimate.[7]

When the uncertainty of the signal depends on the signal magnitude (as in the case of shot noise or proportional noise), then one need only substitute the uncertainty relationship into Equation 5 to obtain the appropriate maximum likelihood expression. For example, when the predominant noise present in a signal arises from fluctuations in measurement sensitivity, such as excitation source flicker, the standard deviation of the signal increases in proportion to signal size. Since the signal and its uncertainty are proportional to g_i, the uncertainty in concentration is constant, independent of i. Substituting this relationship into Equation 5, gives the following maximum likelihood expression for estimating the concentration:

$$\hat{c} = (1/N) \sum_i (z_i/g_i) \qquad (9)$$

For the case of shot noise, where the standard deviation of the signal varies with the square root of the signal magnitude, the maximum likelihood estimate of the sample concentration is given by the peak area.

2.3. Application to Photoacoustic Spectroscopy

Absorption of radiation from a pulsed laser and non-radiative relaxation of the excited states produces a rapid temperature rise in the sample which in turn generates a pressure wave which can be detected by a piezoelectric transducer.[8] Reflections of the acoustic wave within the sample and transducer result in a reproducible high frequency signal which persists for over 50 μs. While the peak compression signal at the start of the wave can be used for quantifying the sample absorbance,[8] the entire acoustic wave carries amplitude information which could be used to provide a more precise determination. The model of the peak shape, g_i, can be obtained from well averaged photoacoustic transients obtained from more concentrated samples.

In order to compute a maximum likelihood estimate the sample absorbance from such data, the relationship between signal errors and signal size must also be determined from replicate measurements. A plot of signal variance versus the square of the signal amplitude from such measurements is generally linear,[6] indicating a strong proportional noise component arising primarily from pulse-to-pulse variation in laser energy. The intercept of this plot is not zero, showing a constant noise source (detector noise) at low signal amplitudes. Since the noise sources are uncorrelated, their variances add so that the overall signal variance is given by:

$$\sigma_{z_i}^2 = k^2 z_i^2 + \sigma_z^2 \qquad (10)$$

where k is the coefficient for proportional noise and σ_z^2 is the constant noise variance. Substituting this mixed proportional and constant noise model into Equation 5 along with Equation 7, gives a maximum likelihood estimate which

weights large signals by $1/g_i$ and small signals by g_i, according to:

$$\hat{c} = \frac{\sum' [z_i g_i / (k^2 z_i^2 + \sigma_z^2)]}{\sum' [g_i^2 / (k^2 z_i^2 + \sigma_z^2)]} \qquad (11)$$

Application of this equation to determining the sample concentration or absorbance from photoacoustic transients is illustrated in Figure 1. The capability of a maximum likelihood estimate to extract quantitative information from noisy data is illustrated by these results. For the 1.9×10^{-5} cm^{-1} absorbance sample in Figure 1b, for example, the maximum likelihood estimate of absorbance is in error by only 20% of the true value, despite the largest peak in the transient being comparable to the noise. The scaled residuals, shown in Figure 1c, are random and of a magnitude expected for the experimental error, indicating that most of the quantitative information has been successfully extracted.

The limit of detection[9] of the maximum likelihood photoacoustic absorbance measurement, determined from replicates, was $A_{min} = 7 \times 10^{-6}$ cm^{-1}, which represents a 5($\pm$2)-fold improvement over single point measurements at the peak of the waveform. This observed result is indistinguishable from the 5.6($\pm$0.2)-fold improvement predicted from a propagation of errors through Equation 11. One can conclude from these results that the method of maximum likelihood, an optimum technique for combining measurements of differing uncertainty, is appropriate for the quantitative interpretation of signal peaks where the peak shape is known in advance or is reproducible and can be measured precisely. Propagation of errors through the maximum likelihood estimate expressions allows one to predict both the uncertainty in the quantitative results and the improvement in precision relative to other methods.

3. ANALYSIS OF ONE-DIMENSIONAL SPECTROSCOPIC DATA (SPECTRA, WAVEFORMS)

3.1. Spectrophotometry of Mixtures

Acquisition of spectroscopic signals as a function of an independent variable (wavelength or time) increases the information content of the measurement[1,2] over zero-dimensional results and allows the identity and composition of complex mixtures to be determined. The linear relationship between absorbance and concentration given by the Beer-Lambert law makes the use of linear regression analysis appropriate for determining the concentration of individual components in the mixture. The measured absorptivity (in cm^{-1}), a_i, of an n component sample at wavelength, i, can be written as the sum of the absorptivity contributions of each component:

$$a_i = k_{i1} c_1 + \cdots k_{ij} c_j + \cdots k_{in} c_n + r_i \quad (12)$$

where k_{ij} is the molar absorptivity of component j at wavelength i, c_j is the molar concentration of component j, and r_i represents the residual error in the measurement. A series of absorbance measurements at m different wavelengths generates a system of m equations, each having the form of Equation 12. It is convenient to express this system of equations in matrix form:

$$\underline{A} = \underline{K}\,\underline{C} + \underline{R} \quad (13)$$

where $\underline{A}$ is a vector of absorbances of the mixture, measured at m different wavelengths (the mixture spectrum), $\underline{K}$ is an m-by-n matrix of standard spectra of the n components measured at each of the m wavelengths, $\underline{C}$ is a vector

containing the n unknown concentrations, and $\underline{R}$ is an m-element vector containing the residual error in the measured absorption spectrum of the mixture. Note that $\underline{K}$ is a model of the wavelength variation which we expect to observe in the mixture spectrum, where each of the columns of $\underline{K}$ is weighted by an element in $\underline{C}$, the concentration of the particular component in the mixture.

Given such a model for mixture spectrum where the concentrations of the components are unknown and an excess of degrees of freedom ($m > n$), one obtains a maximum likelihood estimate of component concentrations by first calculating the sum of the squared residuals with respect to the n concentrations (divided by the constant measurement variance, $\sigma_i^2 = \sigma^2$):

$$\chi^2 = \sum_{i=1}^{m} r_i^2 / \sigma_i^2 \qquad (14)$$

$$= (1/\sigma^2)\, \underline{R}^T \underline{R} \qquad (14a)$$

To obtain a optimize the estimate, this chi-squared statistic[5] is minimized with respect to each of the unknown concentrations by solving a series of n "normal" equations of the form:

$$(\partial \chi^2 / \partial c_j) = 0 \qquad (15)$$

Note that Equation 15 is simply a multivariate form of Equation 4 which was derived for maximum likelihood estimation of a single variable.

This series of n simultaneous equations has a simple linear algebra solution[4,10] given by:

$$\hat{\underline{C}} = (\underline{K}^T \underline{K})^{-1} \underline{K}^T \underline{A} \qquad (16)$$

where the product of the first three terms on the rhs of Equation 16 are often termed a pseudoinverse or least squares inverse of the matrix, $\underline{K}$. A check on the validity of Equation 16 can be made by allowing the residuals to vanish, $\underline{R} = 0$, where the measured absorbance is exactly predicted by the model, $\underline{A} = \underline{K}\,\underline{C}$. Substituting $\underline{K}\,\underline{C}$ for $\underline{A}$ in Equation 16 shows that, under these ideal conditions, the maximum likelihood estimate for the concentrations, $\hat{\underline{C}}$, is identical to the true concentration vector, $\underline{C}$.

3.2. Analysis of Errors in Linear Regression

The uncertainty or variance associated with the concentration estimates extracted from a mixture spectrum can be found by a propagation of errors analysis of Equation 16.[4] The results of such analysis include both the variance of each of the extracted parameters as well as the covariance between parameters. These terms are collected into an n-by-n variance-covariance matrix, $\underline{V}$, with the parameter variances appearing on the diagonal and the covariance terms as the off-diagonal elements. The linear algebra expression for the error analysis has a remarkably simple form,[4] given by:

$$\underline{V} = (K^T K)^{-1} \sigma^2 \qquad (17)$$

where σ^2 is the variance associated with the measured absorbance of the mixture spectrum.

The concentration errors associated with such a spectrophotometric

determination of a mixture arise from the product of two terms on the rhs of Equation 17, which have entirely different origins.[11] The variance term, σ^2, depends only on the precision with which the mixture spectrum is measured, and is thus related to the characteristics of the instrument and experimental methods. The second term, $(\underline{K}^T\underline{K})^{-1}$, is an n-by-n matrix which serves to amplify the measurement error in the estimation of concentrations. The magnitudes of the elements in this matrix, f_{jj}, depend on differences between the spectra of the components and the wavelengths chosen to measure absorbance, the latter being an exercise in experimental design.[11] The greater the similarity between two standard spectra, which appear in the columns of $\underline{K}$, the larger will be the inverse of $(\underline{K}^T\underline{K})$, particularly the diagonal elements corresponding to the similar spectra and the off-diagonal elements between them.

3.3. Selecting "Analytical" Wavelengths

Choosing a set of wavelengths at which to gather absorbance data in order to estimate the concentration of components in a mixture is an historic problem in spectrophotometric analysis.[12-14] Since minimizing the parameter variance (least squares) returns a value of the parameter which maximizes the likelihood of having observed the data, one would optimally design quantitative, spectrophotometric experiments by selecting "analytical" wavelengths which minimize the elements of the variance-covariance matrix. Since the design only affects $(\underline{K}^T\underline{K})^{-1}$ in Equation 16, one need not consider the measurement precision contribution to $\underline{V}$, when selecting wavelengths at which to gather data.

This regression-based concept for selecting "analytical" wavelengths has been evaluated for two-component mixtures.[15] Spectra were modeled as Gaussians as shown in Figure 2, where the distance between the means was varied. To

assess the predicted error in the estimated concentrations as a function of wavelengths in the design matrix, $\underline{K}$, the sum of the diagonal elements of $(\underline{K}^T\underline{K})^{-1}$, $(f_{11} + f_{22})$, is plotted versus the first two wavelengths chosen. The results are shown as a contour plot in Figure 3 for the spectra in Figure 2b. The error surface has equivalent minima at (λ_1,λ_2) = (40,60) and (60,40), where $(f_{11} + f_{22})$ = 3.90. Since the standard spectra are symmetric, $f_{11} = f_{22}$, the minimum value of $(f_{11} + f_{22})$ indicates that replicate determinations of the concentrations of the components by a two-wavelength measurement at the best wavelengths would exhibit a variance which is 1.95 times larger than the variance of the absorbance measurements (see Equation 17). The optimum set of wavelengths represents a distinct minimum, as shown in the slice through the error surface in Figure 2b.

Generating optimal data at more than two wavelengths presents a larger error minimization problem; an effective approach to dealing[15] with this problem is to fix the first two wavelengths at the above values and vary the next pair. The concentration variance which results from this approach is plotted in Figure 4. The results indicate that same pair of wavelengths are optimal for the third and fourth measurements, as for the first two. The concentration variance arising from two measurements each at the same pair of wavelengths is exactly one half that observed for the two wavelength design of Figure 3. This is not surprising since the design has not changed except to double the number of measurements, which improves the measurement variance by a factor two. While the optimal wavelengths in a design remain the same as the number of measurement is increased, the minima in the error surface become much less distinct; this trend is apparent in comparing Figure 3 with Figure 4, and flattening of the error surface continues as m increases. The penalty for

measuring at less than the optimal wavelengths is smaller once measurements at or near the optimum region are included in the design.

To test this trend and its effects on concentration precision, a 50-fold replicated, 2-wavelength experiment design was compared to a measurement of a complete, 100-wavelength spectrum with wavelengths spread uniformly over the range of Figure 2. For the resolution of component spectra R_s = 0.5, 0.25, and 0.1, the concentration variance improved by a factor 3.1, 3.7, and 4.1, respectively, when using the optimal, 2-wavelength design compared to measuring a complete spectrum. The improvement in precision is greatest when the spectra are poorly resolved, and the magnitude of the improvement is rather modest when the number of measurements is large.

This modest gain in precision provided by a replicated n-wavelength design for an n-component determination is offset by a significant penalty: an insensitivity to model error. In the analysis of a mixture spectrum modeled by Equation 13, we have assumed that all of the components in the mixture are represented among the standard spectra in the matrix, $\underline{K}$. If this is not the case, due to an unexpected contamination for example, the vector of residual error, $\underline{R}$, which is the difference between the best fit, $\underline{K}\,\hat{\underline{C}}$ and the measured spectrum, $\underline{A}$, will generally show structure due to the spectral variation not accommodated by the model; furthermore, the magnitude of the residuals will exceed the expected measurement error. In the case of a replicated n-wavelength design for an n-component determination, the residuals are not sensitive to model error and will never exceed the measurement precision. This situation is analogous to fitting calibration data to a straight line by acquiring replicate measurements of the dependent variable at only two points along the x-axis. If these points are at the origin and the extremum of the

x-axis, then the precision of estimating the intercept and slope are optimized. On the other hand, this choice of measurements along the x-axis gives no hint as to whether the data should be fit to a straight-line, that is whether the assumed model is correct. For large numbers of measurements, acquiring data over the entire range of the x-axis returns slightly poorer precision in the estimated slope and intercept, but allows a non-linear response to be detected in the residuals.

3.4. Weighting Observations in One-Dimensional Linear Regression

In the deriving the concentration estimates which maximize the likelihood of having observed a particular mixture spectrum, we have thus far assumed that the measurement variance is constant, as in Equation 14a. The estimated concentrations, $\hat{\underline{C}}$, given by Equation 16 are the values which minimize the sum of the unweighted squared residuals, $\underline{R}^T\underline{R}$. If error of measurement does not satisfy this assumption, then the residuals, $\underline{R}$, are drawn from populations of differing variance, and the $1/\sigma_i^2$ factors in the summation defining chi-square in Equation 14 cannot be equated and brought outside the sum. The normal equations (Equation 15) under these conditions must, therefore, minimize the sum of the squared residuals with each weighted by the inverse of the expected variance; this is analogous to Equation 4 for zero-dimensional data.

A convenient algebraic approach to achieving this goal is to multiply Equation 13 by a weighting factor which makes the elements of the residual vector have the same variance. A factor which will accomplish this goal is an m-by-m diagonal matrix, $\underline{W}$, where the elements, $w_{ii} = 1/\sigma_i$.[4] Multiplying both sides of Equation 13 by this matrix results in the following identity:

$$\underline{W}\,\underline{A} = \underline{W}\,\underline{K}\,\underline{C} + \underline{W}\,\underline{R} \tag{18}$$

The elements of the weighted vector of residuals, ($\underline{W}\ \underline{R}$), are drawn from a population having the same variance (equal to unity). The weighted definition of chi-square has a simple linear algebra form given by:

$$\chi^2 = (\underline{W}\,\underline{R})^T(\underline{W}\,\underline{R}) \tag{19}$$
$$= \underline{R}^T\underline{W}^T\underline{W}\,\underline{R} = \sum_{i=1}^{m} r_i^2/\sigma_i^2$$

where $\underline{W}^T\underline{W}$ is a diagonal matrix whose elements are $1/\sigma_i^2$. Since ($\underline{W}\ \underline{R}$) of *Equation 18 has uniform variance, one can obtain the maximum* likelihood concentration estimates for this equation using the linear algebra solution to the normal equations for the uniform variance case, Equation 16. Multiplying the weighted mixture spectrum, ($\underline{W}\ \underline{A}$), on the left by the pseudoinverse or least squares inverse of ($\underline{W}\ \underline{K}$) gives the concentrations which minimize chi-square of Equation 19:

$$\hat{\underline{C}} = \left((\underline{W}\,\underline{K})^T(\underline{W}\,\underline{K})\right)^{-1}(\underline{W}\,\underline{K})^T(\underline{W}\,\underline{A}) \tag{20}$$
$$= (\underline{K}^T\underline{W}^T\underline{W}\,\underline{K})^{-1}\,\underline{K}^T\underline{W}^T\underline{W}\,\underline{A}$$

The variance-covariance matrix for the estimated concentrations from this weighted least squares analysis arise from the inverse term of Equation 20,

analogous to Equation 17 for the unweighted case:

$$\underline{V} = (\underline{K}^T \underline{W}^T \underline{W} \underline{K})^{-1} \qquad (21)$$

3.5. Application to Time-Resolved Fluorescence Spectroscopy

Fluorescence measurements made with photon-counting detection and stable excitation sources are generally dominated by shot noise, characterized by a Poisson error distribution.[5] As a result, the residual errors arising from fitting spectra or time-decay curves have a variance which is equal to the mean number of counts detected, which for large number of counts (>100) may be approximated (with <20% error) by the number of observed counts, $\sigma_i^2 \sim a_i$. Substituting this approximation into $\underline{W}$, results in the diagonal elements of $\underline{W}^T\underline{W}$ having the value $1/a_i$, so that the product ($\underline{W}^T\underline{W}$ $\underline{A}$) in Equation 21 is an m-element column vector where all elements are unity. Interestingly, all of the information about the measured fluorescence spectrum or time-decay curve used to estimate the concentrations in Equation 21 resides in $\underline{W}^T\underline{W}$, within the inverse.

Recently, this method of data analysis has been applied to quantitative resolution fluorescence decay curves where the lifetimes of the two components are similar.[16] The form of the one-dimensional data is shown in Figure 5 where naphthalene in cyclohexane is repetitively excited with a pulsed laser, and the decay of fluorescence intensity is collected as a time-histogram of single-photon arrivals following the excitation pulse.[17] The decay of the excited state population is governed by first-order kinetics. This single component transient can, therefore, be fit to a single exponential decay of the form:

$a_i = (c_j/\tau_j)\exp(i\Delta t/\tau_j)$, by minimizing chi-square of Equation 19 with respect to c_j and τ_j. While the magnitude of the unweighted residuals show a dependence on signal intensity, as shown in Figure 5a, the residuals weighted by the inverse of the expected shot-noise, shown in Figure 5b, are random and have the same variance.

Time-resolved fluorescence spectroscopy can be used for multicomponent determinations; the technique is especially useful, for example, as an in-situ spectroscopic probe of molecular environments.[18] Since detection of fluorescence intensity is linear, the decay curve of intensity from a multi-component sample can be modeled according to Equations 12 and 13, where $\underline{K}$ contains normalized decay curves of the components, $k_{ij} = (1/\tau_j)\exp(i\Delta t/\tau_j)$, and the vector, $\underline{C}$, contains the total number of photon counts (the pre-exponential factors) which are proportional to concentration. This approach to the quantitative analysis of mixed decay curves allows the use of Equation 20 to efficiently determine the statistically optimal set of pre-exponential factors. If the fluorescence lifetimes of the components are known in advance, then the known vectors which comprise the matrix, $\underline{K}$, can be used to extract concentrations in one step.

The advantage of having a physical model for a spectroscopic process being measured, and thereby a functional form for the data, is that $\underline{K}$ need not be known in advance. By varying the n fluorescence lifetimes, τ_j, defining the vectors in $\underline{K}$, one can determine the particular lifetime values which minimize chi-square according to Equation 19 and thus obtain the best estimate of the matrix $\underline{K}$ for a measured decay curve. While the pre-exponentials and fluorescence lifetimes can both be determined by a search of parameters, more precise parameter estimates are obtained faster by incorporating the linear

least squares determination of $\hat{\underline{C}}$ using Equation 20 into search for the n non-linear parameters, τ_j.[18] A second advantage of using a weighted, linear regression step to estimate the component amplitudes is that the uncertainty of the estimates can be predicted from first principles. Using the relative standard deviations derived from variance-covariance matrix for Equation 20, $\underline{V} = (\underline{K}^T \underline{W}^T \underline{W} \underline{K})^{-1}$, the errors in determining the component amplitudes from a series of measured fluorescence decay curves were predicted and compared with the observed precision found in replicate measurements. As shown in Figure 6, the error predictions of the variance-covariance matrix follow the observed results over a wide range of total photon counts in the data.

4. TWO-DIMENSIONAL SPECTROSCOPIC MEASUREMENTS

4.1. Combinations of Correlated and Uncorrelated Dimensions

The exponential decay of intensity in a time-resolved fluorescence experiment provides an excellent example of a correlated measurement dimension. While fluorescence intensity is measured at hundreds of points in time in such an experiment, the intensity channels are not independent but are related by the functional form of the decay of the components. It is prior knowledge of this relationship which allows the $\underline{K}$ matrix to be determined by fitting only one parameter per component in the sample. While such correlated behavior is valuable for resolving overlapped data from mixtures, the number of degrees of freedom in such a measurement is drastically over data which is less predictable and therefore more informative.[2]

Among the most powerful spectroscopic methods for resolving and identifying components in complex mixtures are "hyphenated"[3] combinations of correlated and uncorrelated measurement dimensions. Examples include GC-MS,

LC-UV, GC-IR, and time-resolved fluorescence spectroscopy. In these methods, the correlated measurement dimension (generally the time-dependence, as in chromatography-spectroscopy combinations) can be used to resolve overlap between the components, either by using a physical model of the response[19-21] or by seeking out correlations in the time-dependence with factor analysis.[22-24] The spectra of components, thus resolved, generally show much richer variation and thus contain more information for identification. In absence of any prior knowledge of what possible components are present, the unpredictable nature of the spectra makes mixture analysis with only this dimension impossible. Therefore, the hyphenated combination of predictable and unpredictable measurement dimensions is ideally suited to determining the composition of a complex sample.

4.2. Modeling the Correlated Dimension: pH - UV Data Analysis

The use of physical models and regression methods for resolving component behavior in a correlated dimension is greatly assisted by measurements taken along a second spectroscopic dimension which is less predictable. Differences in the components along the information-rich spectroscopic variable aid in the convergence of the model. An example of this benefit has been demonstrated for analyzing spectrophotometric titrations by measuring a complete UV-Vis absorption spectrum as a function of pH. A synthetic example of such a data set is shown in Figure 7, where the absorption of mixture of two monoprotic acid/base pairs shows the dominance of the acid forms at low pH and the transformation to base forms at higher pH. Since the absorbances of the components in the mixture are additive, the absorbance at any wavelength, i, and pH, j, is the sum of the contribution of each of the components in the

mixture. The data for a given sample are no longer a single vector but rather a matrix of absorption spectra as a function of pH.

If the spectra of the components are independent of pH where the acid and base forms change only in relative proportion, the response can be modeled according to Equations 12 and 13. The matrix $\underline{A}$ contains the absorption spectra versus pH for the mixture, where i is the index of wavelength down the rows and j is the index of pH across the columns. Like the one-dimensional case, $\underline{K}$ contains the spectra of the n components in its columns, however, $\underline{C}$ is now a matrix containing the pH dependent distribution curves in its rows. For a given measurement of $\underline{A}$, the data analysis task is to decompose the matrix into best estimates $\hat{\underline{K}}$ and $\hat{\underline{C}}$ which is done without advance knowledge of either factor.

To carry out this task, the pH dependent distribution curves will be modeled according to equilibrium theory; to factor the model behavior from the data, however, the rows of $\underline{C}$ must be linearly independent so that a unique, best fit value of $\hat{\underline{K}}$ exists. This can be accomplished by defining the rows of $\underline{C}$ as difference composition curves,[25,26] one for each of the n acid/base pairs in the mixture,

$$c_{kj} = \left[\frac{10^{(pH_i - pK_k)} - 1}{1 + 10^{(pH_i - pK_k)}} \right] \qquad (22)$$

which corresponds to a difference absorption spectrum in $\underline{K}$ given by:

$$k_{ik} = (\epsilon_{A^-} - \epsilon_{HA})_{ik} b ([HA] + [A^-])_k \qquad (23)$$

where the difference in the molar absorptivity of the base and acid forms is

multiplied by the sample path length, b, and the total concentration of the particular acid/base pair. To preserve total intensity in the data, a final, n + 1, row is defined in $\underline{C}$ in which all of the elements are equal to n. This corresponds to an (n + 1) row in $\underline{K}$, which contains the sum of all absorbing species in solution including non-pH varying species.

For a given estimate of the pH dependent composition, $\hat{\underline{C}}$, which requires an estimating the n pK_a's, the least squares set of difference spectra can be found by multiplying the data matrix, $\underline{A}$, by the right-pseudoinverse of $\hat{\underline{C}}$, which is analogous to Equation 16 above,

$$\hat{\underline{K}} = \underline{A}\,\hat{\underline{C}}^T(\hat{\underline{C}}\,\hat{\underline{C}}^T)^{-1} \qquad (24)$$

The quality of the fit of the product ($\hat{\underline{K}}$ $\hat{\underline{C}}$) to the data depends on the accuracy of the estimated pK_a's defining $\hat{\underline{C}}$. To test the quality of fit, the value of chi-square (Equation 14 for the unweighted case) is calculated. The estimates of the pK_a's are varied so as to minimize chi-square, usually by a Nelder-Mead simplex algorithm[27,28]; at each step of the non-linear least squares search for the pK_a's, the linear least squares step of Equation 24 returns the best estimate of the matrix $\underline{K}$ for a given estimate of $\underline{C}$.

4.3. Acid/Base Mixture Resolution and Error Predictions

To test this method of resolving mixtures of monoprotic acids, data matrices containing absorption spectra of mixtures of two and four acid/base indicators were acquired at intervals of 0.2 pH units over a pH range of 3.0 to 8.4. A plot of the four component data matrix is shown in Figure 8; the composition of the sample is listed in Table I. From the shape of the data surface, the spectral and pH variation of three acid/base components is

apparent, but behavior of a fourth component is not obvious. It is, however, clear that the component spectra and pH dependences are severely overlapped. Despite the severity of the overlap, reiterative application of Equation 24 to determine the least squares difference spectra as the n pK_a's are varied to minimize chi-square results in an optimal fit to the data matrix, with the results summarized in Table I. The accuracy of the difference spectra which are extracted from the mixture is illustrated in Figure 9 where the results are compared with spectra of the individual components. The quality of fit is good without any systematic error. The differences between the pK_a's determined from fitting the mixture and those obtained by fitting the isolated indicators average 0.07 pH units, which is much less than the 0.2 pH interval between spectral scans in the data.

The accuracy of the resolved difference spectra extracted from the mixture is again predictable from first principles. The right-pseudoinverse in Equation 24, $(\underline{C}^T(\underline{C}\ \underline{C}^T)^{-1})$, contains (n + 1) columns corresponding to the (n + 1) columns of $\hat{\underline{K}}$. These column vectors are multiplied by the corresponding rows of the mixture absorbance matrix to extract the estimated spectra in $\hat{\underline{K}}$, one row at a time. As a result, the error of this least squares solution depends only on the variance of the mixture absorbance and the degree to which the rows of $\underline{C}$ are overlapped, neither of which depend on wavelength. We can, therefore, use the variance-covariance matrix of Equation 17, which for the right-pseudo-inverse is $\underline{V} = (\underline{C}\ \underline{C}^T)^{-1}\sigma^2$, to predict the magnitude of the error in the absorption spectra. Taking a value for the absorbance error from the root mean squared residuals, $\sigma = 5.5 \times 10^{-3}$ a.u., the standard deviations of estimating the difference spectra are predicted and listed alongside the observed error in Table I. The agreement between the predicted and observed errors is reassuring.

5. CONCLUSIONS

Regression methods of spectroscopic data analysis have their theoretical roots in the method of maximum likelihood and least squares analysis. This theoretical basis allows these methods to be derived from simple statistical concepts and applied to multidimensional data. The use of regression method with models takes advantage of correlations in data in order to reduce the effect of noise and resolve overlapped component responses. Without exploiting the prior information available in such correlations, they would only reduce the information content of the measurement without providing any benefit. A particular advantage of modeling is realized with hyphenated spectroscopic methods which combine correlated and uncorrelated measurement dimensions. Modeling the response in the correlated dimension resolves overlap of multi-component samples and provides a pure component response in the uncorrelated, more informing dimension. Using the theoretical basis of regression allows one to predict, a priori, the errors of a data analysis procedure. This analysis of errors can be used to properly weight observations contributing different uncertainty to a result, to optimally design an experiment in terms of what observations are made, and to know in advance what errors to expect in the result.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation through grant CHE85-06667 and by the Office of Naval Research. Fellowship support to J.M.H. from the Alfred P. Sloan Foundation is also acknowledged.

REFERENCES

1. H. Kaiser, Anal. Chem. 42(2), 24A (1970).
2. T. Hirschfeld, Anal. Chem. 48(1), 16A (1976).
3. T. Hirschfeld, Anal. Chem. 52, 297A (1980).
4. N. R. Draper and H. Smith, Applied Regression Analysis, Wiley, New York, (1981).
5. P. R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, (1969).
6. P. E. Poston and J. M. Harris, Anal. Chem. 59, 1620 (1987).
7. R. B. Lam, Appl. Spectrosc. 37, 567 (1983).
8. A. C. Tam and C. K. N. Patel, Appl. Opt. 18, 3348 (1979).
9. L. A. Currie, Anal. Chem. 40, 586 (1968).
10. G. Strang, Applied Linear Algebra, Academic Press, New York, (1976).
11. S. N. Deming and S. L. Morgan, Experimental design: a Chemometric Approach, Elsevier, New York, (1987).
12. F. P. Zscheile, H. C. Murray, G. A. Baker, and R. G. Peddicord, Anal. Chem. 34, 1776 (1962).
13. Z. Przybylski, Chem. Anal. (Warsaw) 14, 1047 (1969).
14. J. Sustek, Anal. Chem. 46, 1676 (1974).
15. S. D. Frans and J. M. Harris, Anal. Chem. 57, 2680 (1985).
16. A. L. Wong and J. M. Harris, Anal. Chem. (submitted).
17. D. V. O'Connor and D. Phillips, Time-Correlated Single Photon Counting, Academic Press, New York, (1984); Chapter 6.
18. J. R. Lakowicz, Principles of Fluorescence Spectroscopy, Plenum, New York, (1983).
19. F. J. Knorr and J. M. Harris, Anal. Chem. 53, 272 (1981).

20. F. J. Knorr, H. R. Thorsheim, and J. M. Harris, Anal. Chem. 53, 821 (1981).

21. S. D. Frans, M. L. McConnell, and J. M. Harris, Anal. Chem. 57, 1552 (1985).

22. W. H. Lawton and E. A. Sylvestre, Technometrics 13, 617 (1971).

23. M. A. Sharaf and B. R. Kowalski, Anal. Chem. 53, 518 (1981).

24. P. J. Gemperline, "Factor Analysis of Spectro-Chromatographic Data", a chapter in this volume.

25. R. I. Shrager and R. W. Hendler, Anal. Chem. 54, 1147 (1982).

26. S. D. Frans and J. M. Harris, Anal. Chem. 56, 466 (1984).

27. J. A. Nelder and R. Mead, Comput. J. 7, 308 (1965).

28. S. N. Deming and S. L. Morgan, Anal. Chem. 45, 278A (1974).

Table I. Reiterative Least Squares Resolution of Acid/ Base Mixtures

Sample Composition	pK_a isolated	pK_a mixture	Observed Spectral Error, s_k	Predicted Spectral Error, σ_k*
methyl orange	3.27	3.50	5.4×10^{-3}	5.2×10^{-3}
bromcresol green	4.77	4.74	3.2×10^{-3}	2.3×10^{-3}
methyl orange	3.27	3.39	5.5×10^{-3}	5.8×10^{-3}
bromcresol green	4.77	4.80	3.8×10^{-3}	4.4×10^{-3}
chlorophenol red	6.07	6.05	3.4×10^{-3}	3.9×10^{-3}
phenol red	7.71	7.71	4.4×10^{-3}	3.7×10^{-3}

* From variance-covariance matrix

FIGURE CAPTIONS

1. Maximum likelihood absorbance estimates from photoacoustic signals.[6] Photoacoustic transients from azulene in carbon tetrachloride in parts a and b have absorbances, A = 4.7 x 10^{-5} cm^{-1} and 1.9 x 10^{-5} cm^{-1}, respectively. The best fits to the data ($\hat{c}g_i$) are shown as heavy lines. Part c show the residual error for part b, scaled by the expected error.

2. Model spectra for two-component mixtures: (a) λ_{max} = 40, 60; resolution, R_s = 0.5. (b) λ_{max} = 45, 55; resolution, R_s = 0.25. (c) λ_{max} = 48, 52; resolution, R_s = 0.1. The width (standard deviation) of the Gaussian peaks is 10 wavelength units. R_s is the difference in the means of the two Gaussian peaks divided by 4-times the standard deviation.

3. Variance in concentration estimates versus two analytical wavelengths for spectra from Figure 2b. (a) Contour plot of (f_{11} + f_{22}); increment between contours is 1.73. (b) Horizontal slice through (a) at λ_1 = 40.

4. Variance in concentration estimates for a four-wavelength design. Wavelengths (40,60) are preselected according to Figure 3, and the next two wavelengths are varied. Minima at (40,60) and (60,40) are indicated, where (f_{11} + f_{22}) = 1.95. (a) Contour plot of (f_{11} + f_{22}). (b) Horizontal slice at λ_3 = 40.

5. Fluorescence decay curve for naphthalene in cyclohexane. (a) The data are fit to a single exponential decay; (b) the residuals are weighted by the predicted error, $r_i/(a_i)^{1/2}$.

6. Observed and predicted errors in quantitative analysis of two-component fluorescence decay curves. The lifetimes are τ_1 = 107.7 ns and τ_2 = 85.0 ns. Solid squares are the relative standard deviation (rsd) predicted for the shorter-lived component using the variance-covariance matrix; circled symbols

are the observed rsd for this component. The x-axis indicates the total photons counted in the measurement, $(a_1 + a_2)$.

7. Synthetic data for a spectrophotometric titration of a binary mixture of monoprotic acids.

8. Spectrophotometric titration of four acid/base indicators. See Table I for sample composition.

9. Difference acid/base spectra (solid lines) resolved from the four component sample of Figure 8. PR is phenol red; CR is chorophenol red; BG is bromcresol green, and MO is methyl orange. Dashed lines are the spectra of the individual components, plotted for comparison.

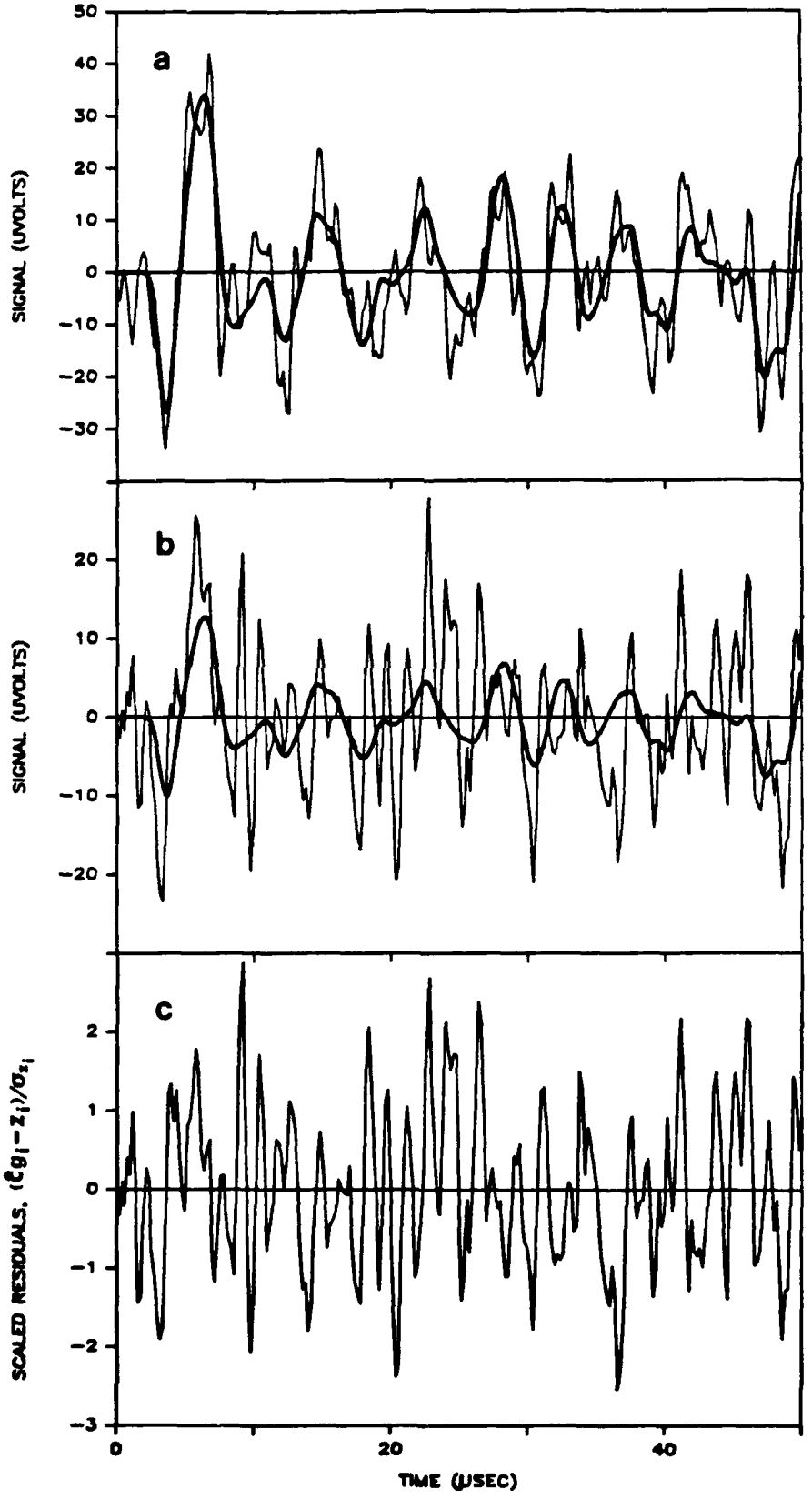

a
b
c
SIGNAL (UVOLTS)
SIGNAL (UVOLTS)
SCALED RESIDUALS, $(\ell_{g_i} - z_i)/\sigma_{z_i}$
TIME (USEC)
50
40
30
20
10
0
-10
-20
-30
20
10
0
-10
-20
2
1
0
-1
-2
-3
0
20
40

Fig 1

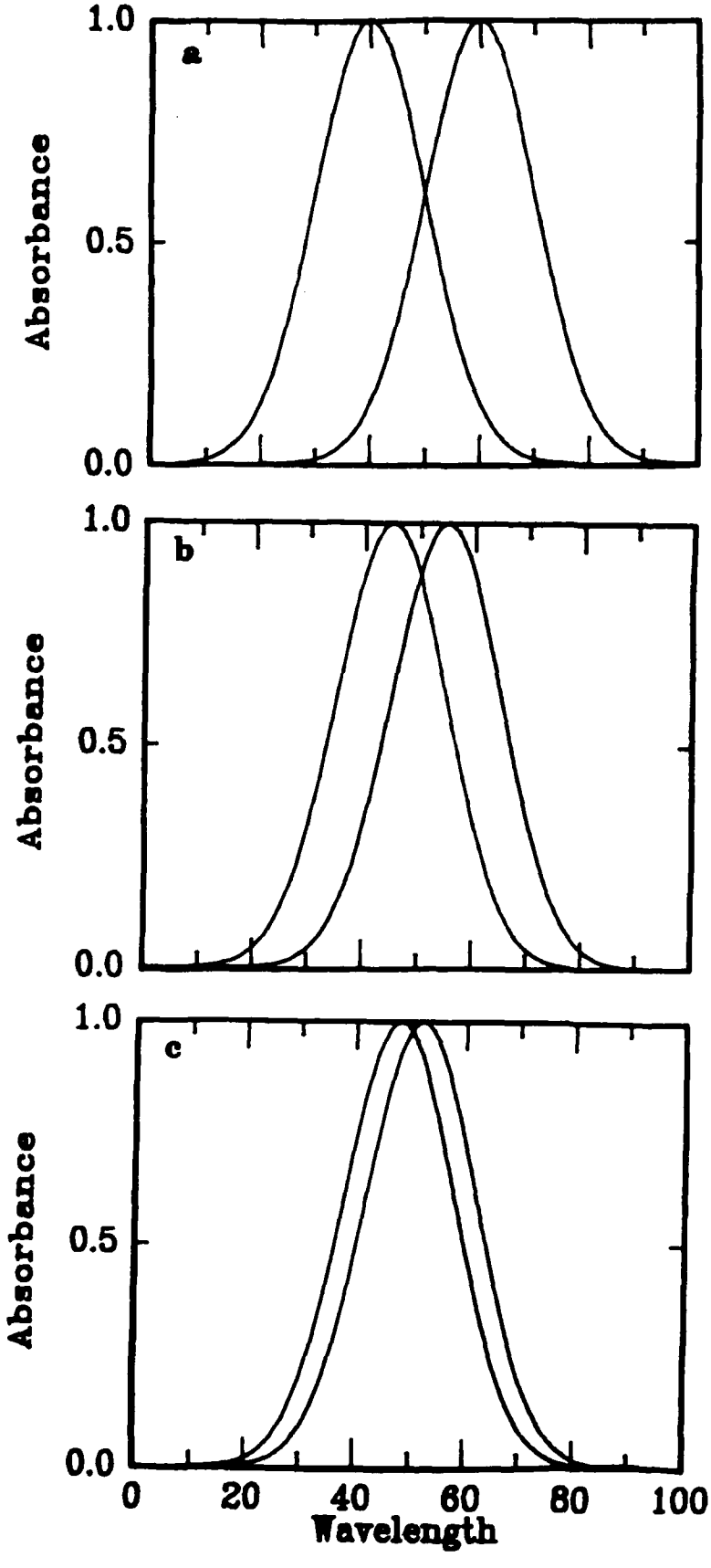


Fig 2

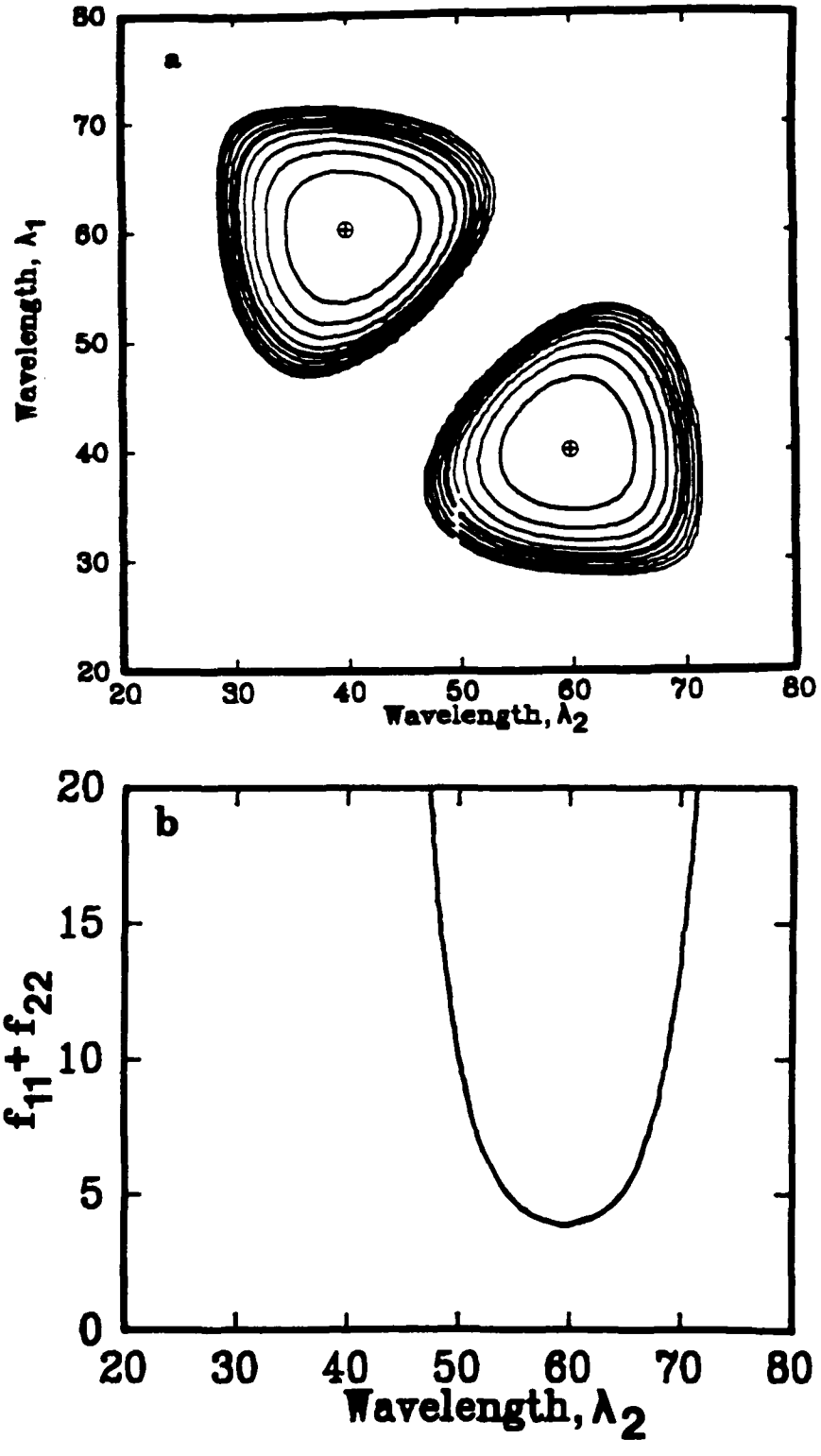
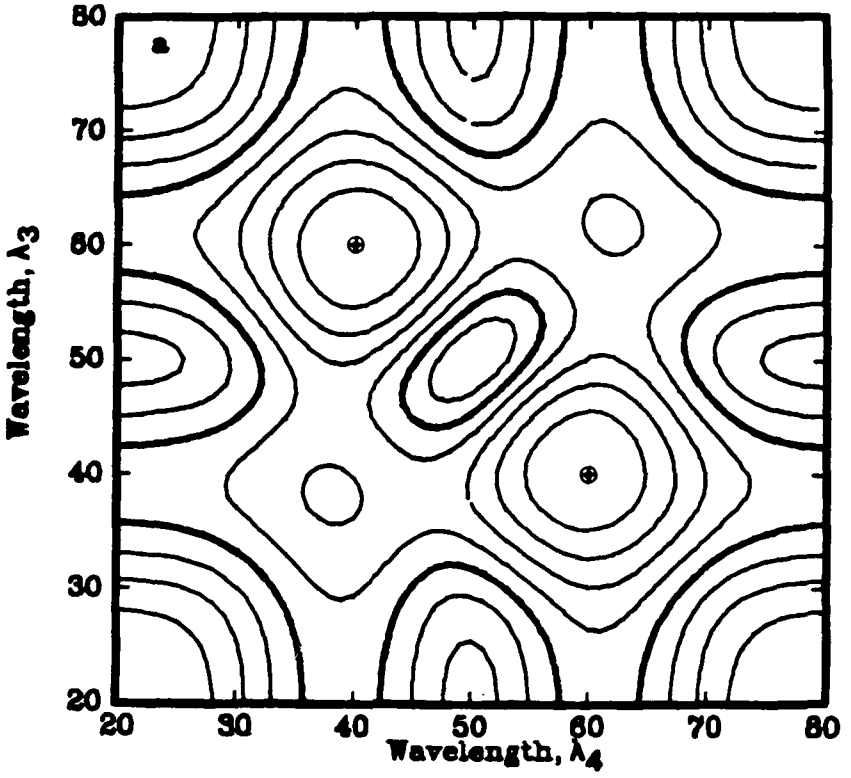

a
Wavelength, λ_1
Wavelength, λ_2
b
$f_{11} + f_{22}$
Wavelength, λ_2

Fig 3


a
80
70
60
50
40
30
20
Wavelength, λ3
20
30
40
50
60
70
80
Wavelength, λ4

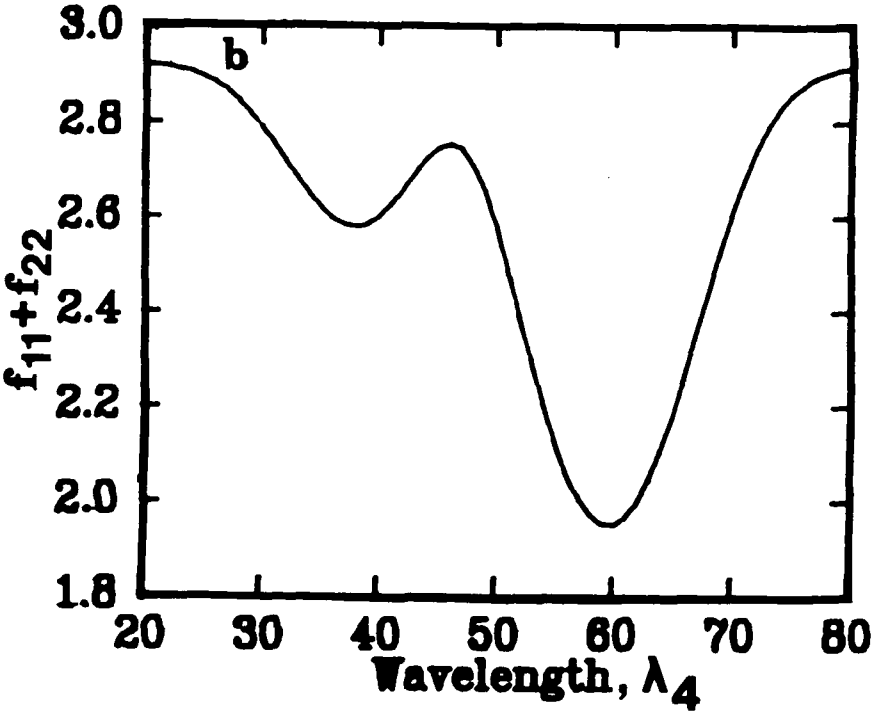

b
3.0
2.8
2.6
2.4
2.2
2.0
1.8
f11+f22
20
30
40
50
60
70
80
Wavelength, λ4

Fig 4

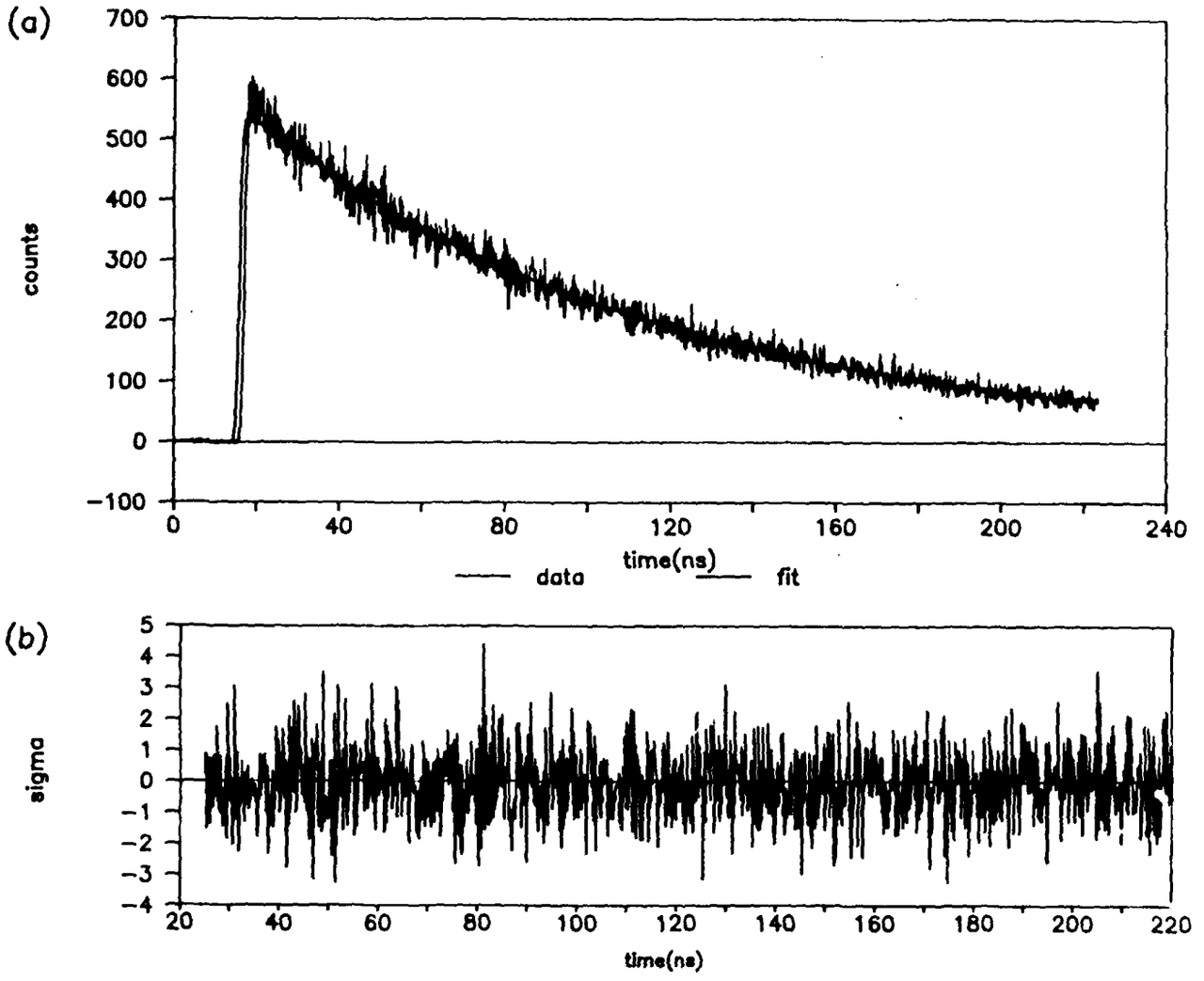

(a)
700
600
500
400
300
200
100
0
−100
counts
0
40
80
120
160
200
240
time(ns)
data
fit
(b)
5
4
3
2
1
0
−1
−2
−3
−4
sigma
20
40
60
80
100
120
140
160
180
200
220
time(ns)

Fig. 5

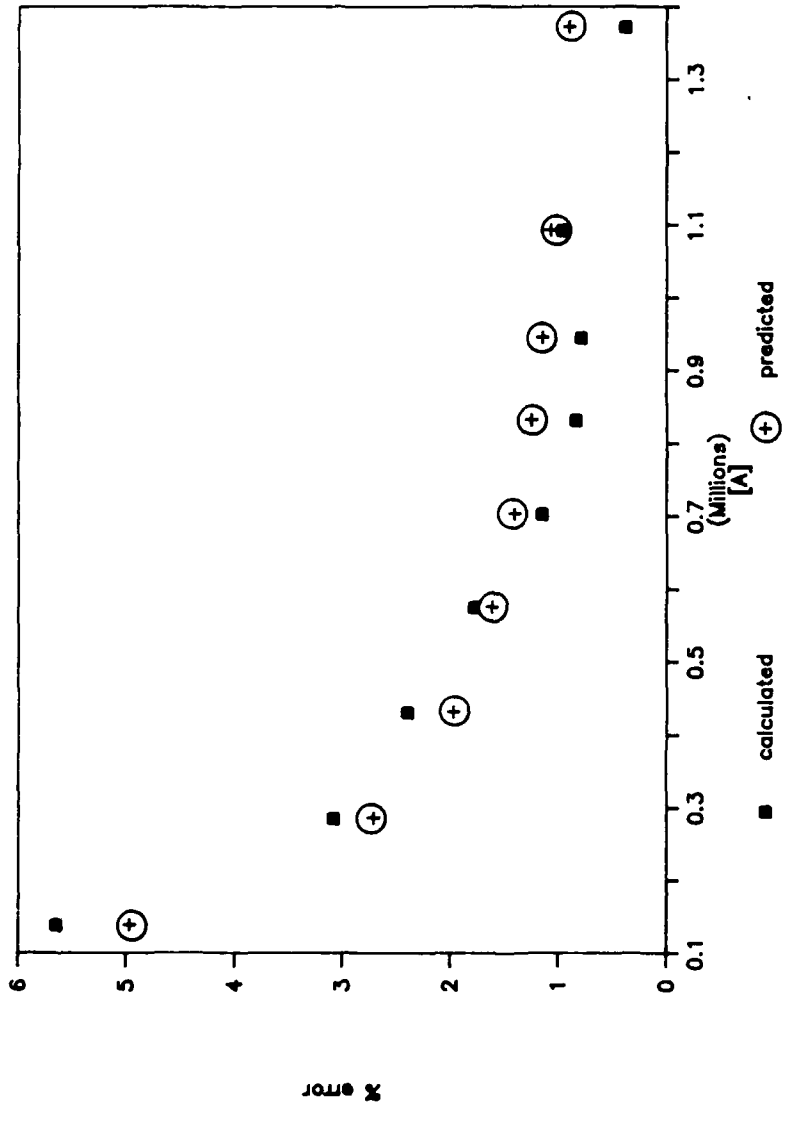

% error
6
5
4
3
2
1
0
0.1
0.3
0.5
0.7
0.9
1.1
1.3
(Millions)
[A]
calculated
predicted

Fig 6

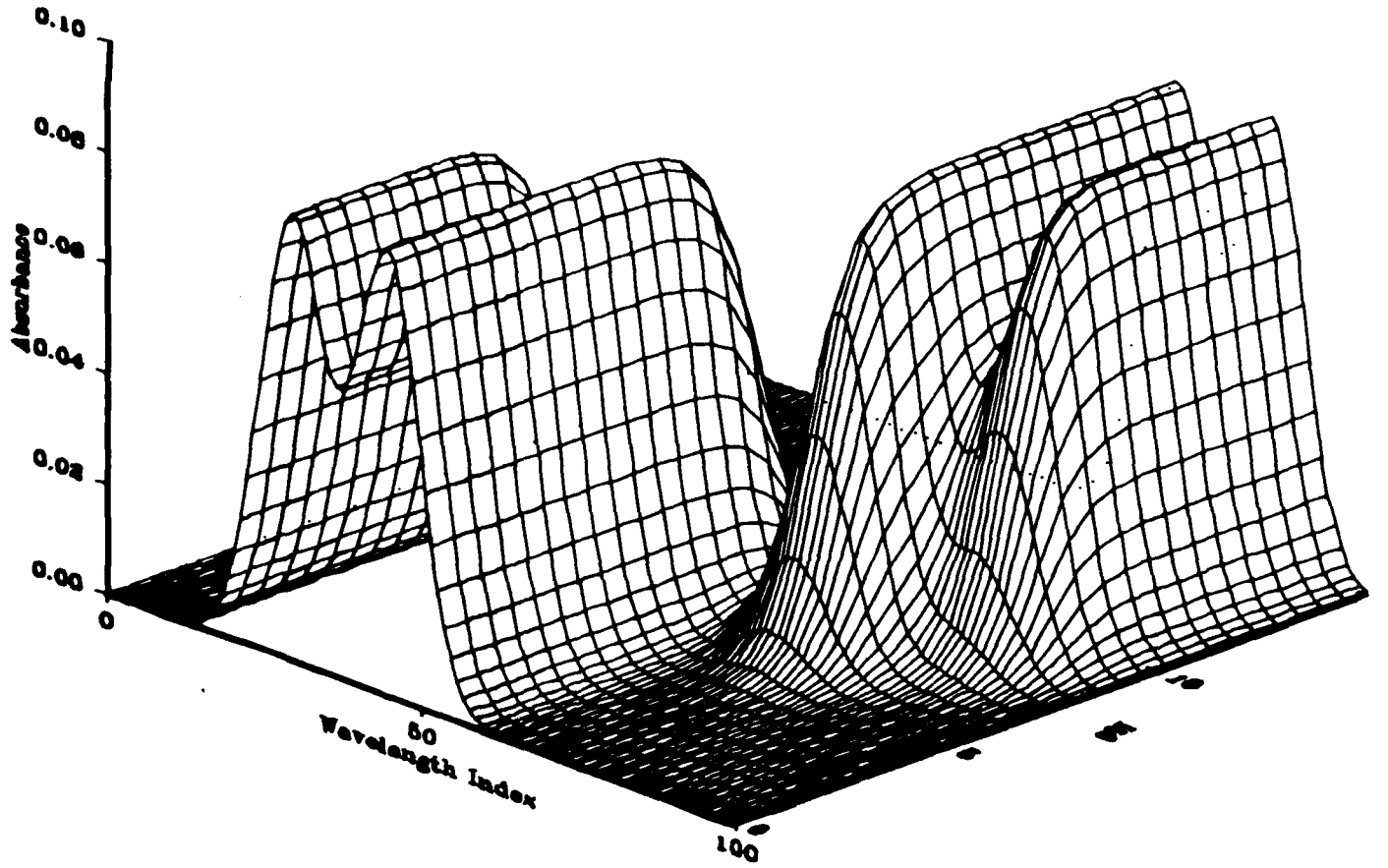


Fig 7

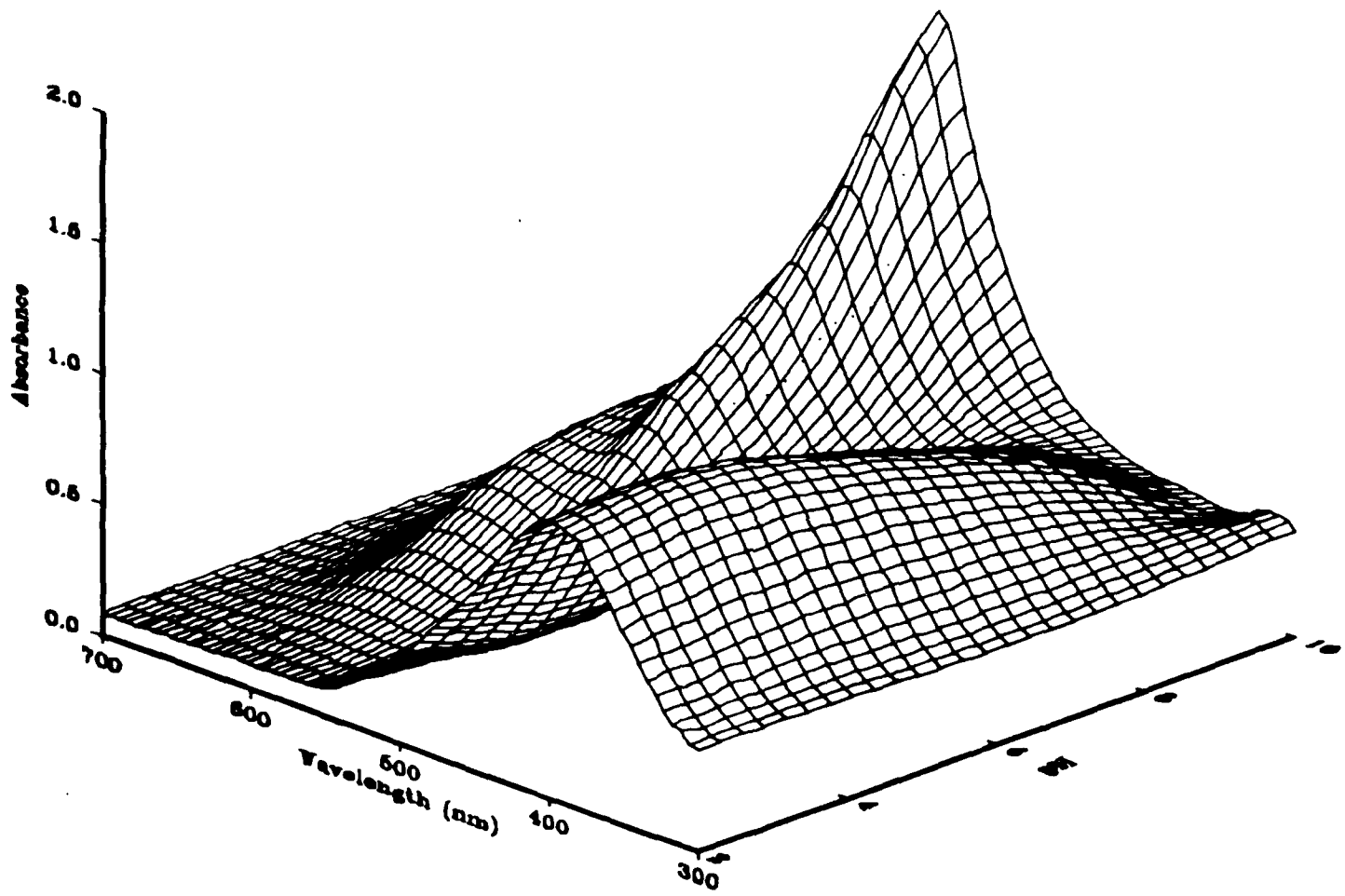


Fig 8

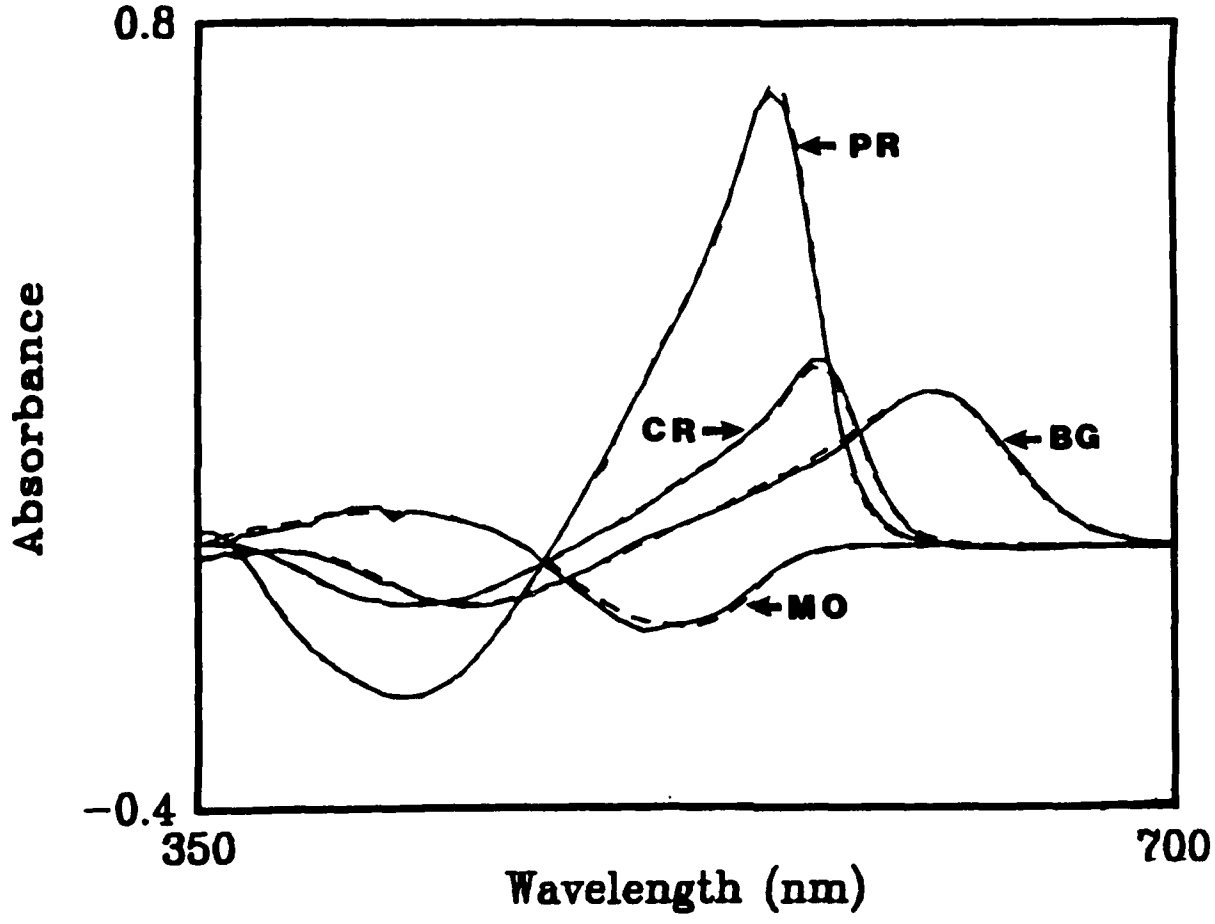


Fig 9